Intel Accelerates CMCC Cloud for 1M IOPS Cloud Disk with Ultra Cloud Storage Experience

INTRODUCTION

Nowadays, with the explosive growth of data caused by the intersection of IoT, AI and 5G related technologies, enterprises are facing higher and higher cost of hyperscale data's storage and management. While along with migrating databases, real-time log analysis and other critical business onto the cloud, enterprises also have higher expectation for storage system's performance. The storage system faces the dual challenge of providing the maximum performance as well as managing the bulk amount of data. As one of the leading Cloud Service Providers in China, China Mobile Communication Corporation (CMCC) Cloud collaborates with Intel to achieve higher storage system performance. The two companies co-created CMCC Ultra Speed Cloud Disk which is capable of millions of IOPS for ultra-high IO performance, 4GB/s of high bandwidth, hundred microseconds of ultra-low latency, and PB layer of system scalability. The new Ultra Speed Cloud Disk fully meets users' demands for outstanding storage performance, which is extreme high IOPS, high bandwidth, low latency, and high scalability, in the Data Technology era.

CHALLENGE: NVMe SSD Requires Faster IO Processing

The new era for NVMe SSD has arrived. With the advantage of its ultra-low latency and IO parallel processing capabilities, the NVMe interface is gradually replacing SAS and SATA, which have been the mainstream storage interconnects in the past 20 years. NVMe SSDs are becoming prominent in market.

For the traditional data storage and processing methods such as SATA/SAS HDD/SSD, each IO needs to be "Interrupted" to transfer data frequently between the user space and the kernel space, where the whole processing requires multiple CPU context switches and memory copies of data. These traditional processing methods are too outdated and inefficient to fully utilize the benefits of NVMe SSDs which are low latency and high concurrency. For this reason, Intel has developed a High-Performance Storage Kit, SPDK (Figure 1).

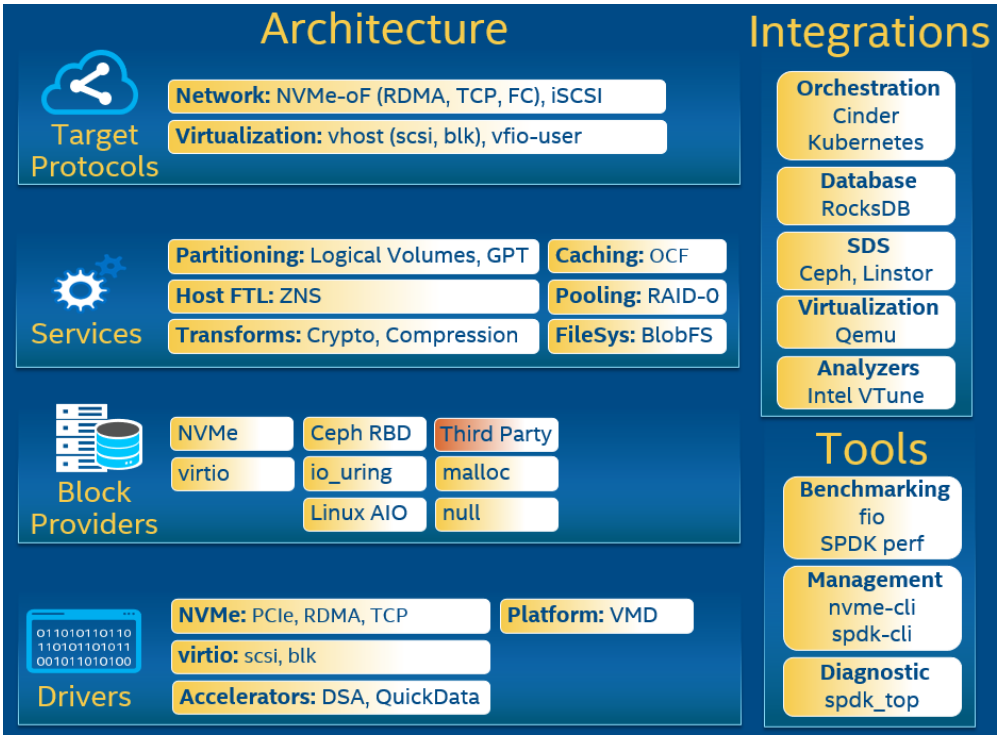

Figure 1. SPDK Architecture Diagram

Intel SPDK architecture majorly consists of four layers:

- **Applications and Protocols Layer:** This layer includes SPDK supported storage applications with corresponding protocols. SPDK iSCSI Target provides standard

iSCSI service for external storage, and users can use the host running through iSCSI host service as a standard iSCSI storage device; SPDK vhost-scsi, vhost-blk or vfio-user provide QEMU with back-end storage services, based on which QEMU can mount virtio-scsi, virtio-blk or NVMe disks for virtual machines; SPDK NVMe-oF provides storage service based on NVMe network transports like RDMA, TCP and so on for external storage.

- **Storage Services Layer:** This layer achieves the abstraction of block, object and file interfaces to offer more storage services. Currently it has implemented the features of QoS (Quality of Service), block data compression and encryption, block level cache, logic volume management and other functionalities.
- **Block Devices Layer:** This layer provides uniform block device services to support different backend storage devices and services, including local and remote storage devices, high-performance NVMe SSDs, AIO devices, Ceph RBD devices, virtio devices and so on. It also enables customized 3rd party block storage devices in SPDK's unified block device management.
- **Drivers Layer:** This layer achieves high-performance user space driver to support different storage devices and hardware accelerations, including local PCIe NVMe driver, network transport of RDMA, TCP for NVMe driver, virtio-blk and virtio-scsi drivers for virtualization. Meanwhile, it also supports hardware devices such as QuickData, DSA and VMD.

According to above support from Intel SPDK Applications and Protocols, Intel SPDK provides back-end storage service for QEMU through vhost-blk. The following shows the analysis of SPDK IO stack model, using the Intel SPDK front-end configuration of vhost-blk, and back-end configuration of NVMe SSDs (Figure 2[1]).

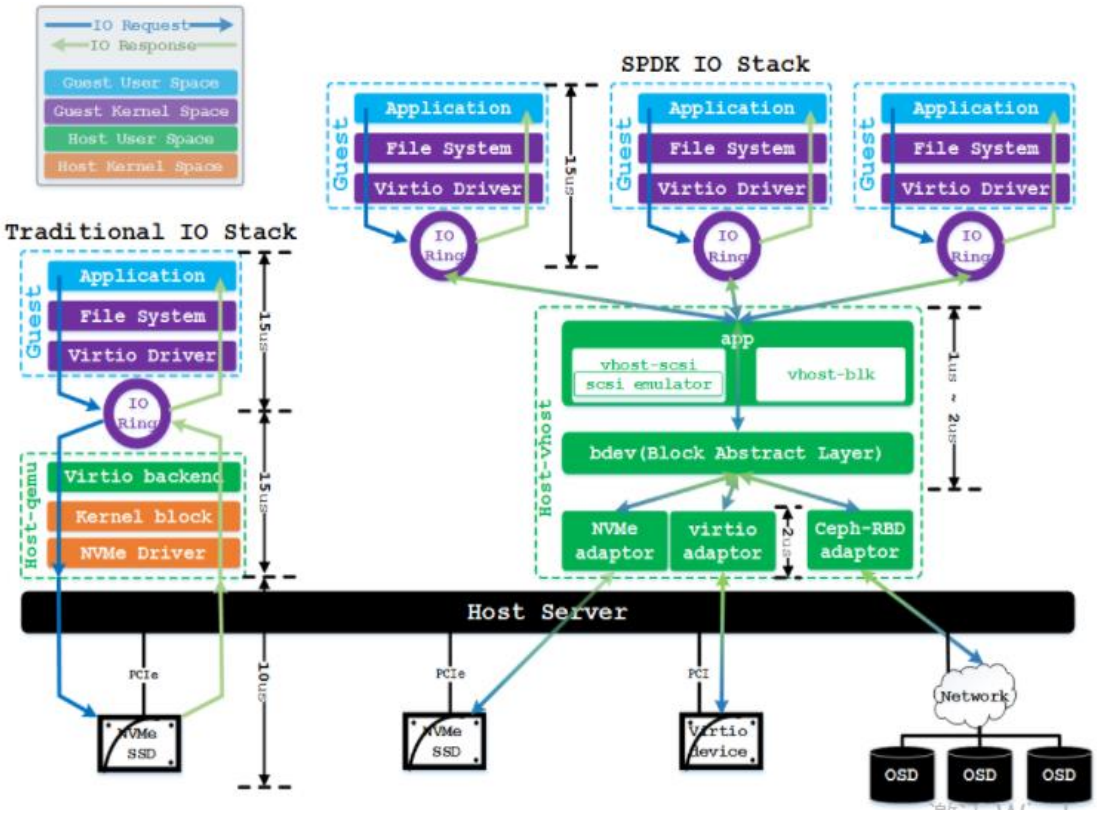

Figure 2. Traditional and SPDK IO Stack Differences

Reference 1：https://rootw.github.io/2018/05/SPDK-iostack/

Virtual machine's processing differences for IO requests:

- **IO Submission Difference:** Traditional NVMe SSDs driver pass requests to the kernel block layer and the kernel NVMe driver for processing via system calls. However, SPDK vhost service continuously extracts requests from the IO ring by polling (meaning that the virtual machine does not have to notify the virtual device when it submits IO requests). For each request extracted, vhost sends it to the Bdev abstraction layer for processing as a task.
- **IO Response Difference:** Traditional kernel NVMe driver notify the QEMU IO thread with a physical interrupt after the NVMe SSD controller completes processing which is also through the interrupt notification. Then the traditional processing puts the response into the VM IO ring and informs the VM of the completion of the request with a virtual interrupt. Nevertheless, SPDK vhost service with the user space NVMe driver polls the Queue Pair of the physical NVMe device. If there is a response, this processing will be

completed immediately without waiting for a physical interrupt from the NVMe SSD device.

- **IO Thread Model Difference:** Traditional kernel NVMe driver uses the multi-threaded and lock-based processing model to operate front-end virtual machine IO Ring or back-end IO Channel, while each IO ring or IO channel of SPDK will only be polled in only one vhost thread as the Run-To-Completion optimization. SPDK vhost can avoid concurrent operation of the same object by multiple threads and allow lock-free method to operate the IO ring or IO channel.

Intel SPDK achieves the great performance optimization of IO latency and throughput through IO polling and lock-free processing. After the end-to-end testing, the results show as follows.

- **IO Latency:** The total latency of the traditional NVMe IO stack is about 40us, while the SPDK user space NVMe IO stack latency is less than 30us. The latency is more than 25% lower (as shown in Figure 2).
- **Throughput:** Traditional NVMe IO stack in a single QEMU IO thread processing can achieve up to 200K IOPS, while SPDK vhost can reach 1 million IOPS when processed by a single thread. With the same CPU overhead, the throughput is 5 times higher by SPDK.

Utilizing the great advantage of IO low latency and high throughput of Intel SPDK, the Cloud Storage team of CMCC Cloud and Intel SPDK co-designed a 1M IOPS Cloud Disk with ultra-high storage performance through the in-depth cooperation.

SOLUTION: With SPDK + RMDA to Build 1M IOPS Cloud Disk

The overall architecture of CMCC Ultra Speed Cloud Disk (as shown in Figure 3) includes Block Storage Interface, Block Storage Service, and Unified Storage Engine.

- **Block Storage Interface:** By utilizing Intel SPDK technologies, this stage implements the EBSdriver, which is a customized client driver for cloud disks. It is mainly responsible for receiving IO from KVM+QEMU and distributing IO requests to different storage nodes through the Block Storage Service based on cloud storage routing information.
- **Block Storage Service:** Blockfs is responsible for block data index management, cloud IO distribution management and other functions.
- **Unified Storage Engine:** Megrez focuses on the data plane. It is able to provide large-scale horizontal scalability of data clusters, cluster disaster tolerance and other service governance capabilities. It also supports functions such as data reliability management.

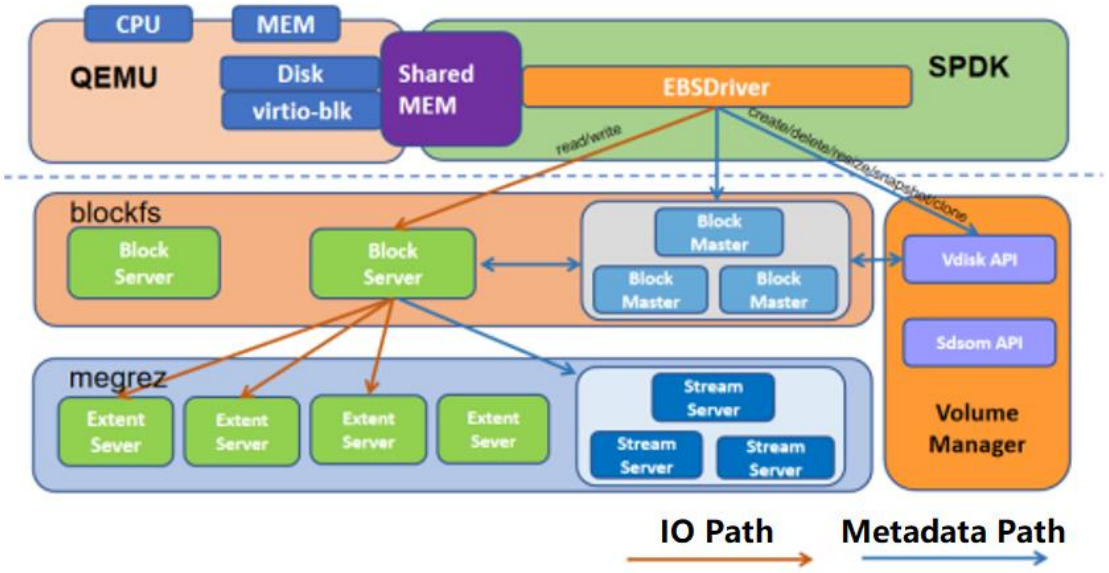

Figure 3. Ultra Speed Cloud Disk Architecture Diagram

The CMCC Ultra Speed Cloud Disk has achieved four innovative technologies in the architecture design.

➢ Centralized Metadata Design for Faster Respond to IO Requests

One of the design concepts of the CMCC Ultra Speed Cloud Disk is to support hyperscale clusters. To achieve hyperscale cluster governance, it is first required to figure out the index scale when the data volume increases, which has long been discussed among the industry. With many years' experience, study and research in the distributed storage domain, CMCC

Cloud Storage Team selected to use the index of centralized and shared metadata design instead of share-nothing design. It is because share-nothing design makes cluster management and capacity scalability more complicated. On the contrary, centralized index supported the distributed storage system with the advantages of flexible scheduling and simple architecture.

The scale of the number of index data is reduced by increasing the granularity of data storage. To achieve this function, the data management unit of CMCC Ultra Speed Cloud Disk has a storage capacity of GB-level slice and can respond quickly to metadata access requests via in-memory storage index metadata. In this way, a set of common x86 servers of metadata index clusters can be used to meet the demand of a single >100PB storage cluster.

- **Implement Write Append Property and Support Non-stop Write**

The centralized index management and write append data model enable fast, non-stop selection and replacement of available data nodes. When a node is sensed to be unavailable, a new set of replica nodes can be immediately allocated to append data write for the subsequent IOs. Then the IO non-stop write feature can be implemented by just updating the new allocated data blocks pointed by index. This feature effectively eliminates IO jitter caused by network anomalies and secures low latency and high throughput of the storage system.

- **Implement Copygroup Copyset, Reduce Frequency of Replica Data Loss**

With the data and storage clusters scale increase, multi-copy technology is facing the risk of data loss when encountering data center associated failures (large-scale clusters reboot or power failure), which arouses a growing concern in the industry. To address this problem, CMCC Ultra Speed Cloud Disk was developed according to the impact of multiple models of distributed storage system using copy modes for the reliability of the storage system in the industry (shown in Figure 4, *Copysets: Reducing the Frequency of Data Loss in Cloud*

Storage). Finally, the replica management capability based on CopySets was designed and added in the Cloud Disk System. By dividing all nodes into N/R CopySets [N is the number of nodes, R is the number of copies], the CopySet is evenly selected based on the pre-defined policy when writing data, so as to effectively reduce the risk of data loss.

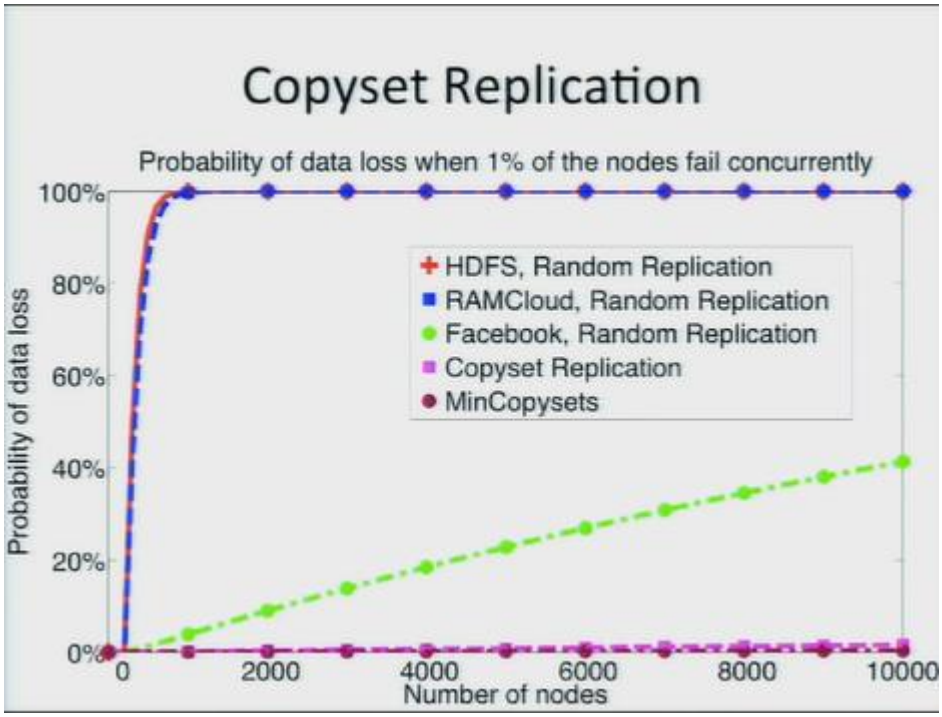

Figure 4: CopySet Replication for the Ultra Speed Cloud Disk

➢ RDMA+SPDK+CPU affinity to Secure the 1M IOPS

CMCC Ultra Speed Cloud Disk added the support of Remote Direct Data Access (RDMA) protocol (as shown in Figure 5) to optimize the server side latency of data processing from the network transfers. RDMA can quickly transfer data from one system to the remote memory without affecting the operating system. It not only eliminates the overhead of data replication and process context switching, but also frees memory consumption, bandwidth, and CPU processing.

Intel SPDK supports polling-mode, asynchronous, lock-free NVMe driver to provide zero-copy, highly parallel capability for direct access to NVMe SSDs from user space applications.

CPU affinity ensures that the single application can only be dispatched to the single CPU core. It effectively eliminates the performance impact of CPU processing from being hampered, such as cache misses, etc., and can lead to more efficient processing and better performance.

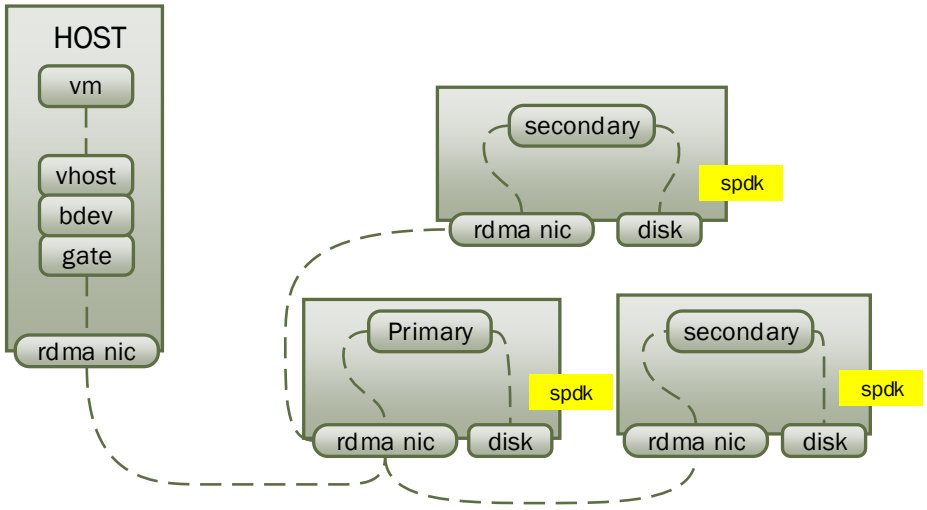

Figure 5. RDMA + SPDK Integrated

RESULTS: Faster and Smarter: 1M IOPS Era by CMCC Ultra Speed Cloud Disk

By making full use of the above advanced architectural design and continuous technology improvement, CMCC Ultra Speed Cloud Disk excels in performance, scalability, protocol support, etc.

- **Extreme Performance in Read and Write:** Based on high-performance Intel NVMe SSDs and CMCC Cloud self-developed unified storage engine, CMCC Ultra Speed Cloud Disk can provide users with 1 million random IOPS per cloud disk, up to 4GB/s throughput, 500us average system latency, and 200us average single IO path latency to achieve the extreme performance for applications.

- **Flexible Clusters Scalability:** CMCC Ultra Speed Cloud Disk allows users to freely configure storage capacity and expand capacity based on their business requirements. The current system can support up to 32TB of a single cloud disk capacity to meet the demand of ultra-large storage space.
- **Various Protocols Support:** Supports Vhost, NVMe-oF protocols for virtualization, bare metal, containers, and other application scenarios.

Figure 6. CMCC Ultra Speed Cloud Disk Capabilities

LOOKING FORWARD: In-depth Cooperation with Intel to Provide Users with Ultra Cloud Storage Experience

CMCC Ultra Speed Cloud Disk represents China Mobile Cloud Center Cloud Storage team's great achievement in the cloud storage domain after years of effort. It is a brand-new self-developed distributed storage solution which satisfies current business requirements and aligns with the development trend in the storage field. With the new 3rd generation Intel® Xeon® scalable processors which have more processing cores, optimized architecture design and larger memory capacity, it will help CMCC Ultra Speed Cloud Disk achieve an even stronger performance. Meanwhile, CMCC Cloud will continuously takes the advantage of new Intel OPTANE SSD and/or non-volatile memory OPTANE PMEM, as well as RDMA NIC in the following in-depth collaborations to achieve more performance improvement.

Looking forward, CMCC and Intel will keep working together on the cloud disk evolution by using the latest technologies and benefit our cloud storage users. We will both strive to

accelerate business and applications on cloud, enable users to achieve cloud service and usage with confidence, pride and best experience.